To appear in IEEE Privacy & Security 15(6), Nov./Dec. 2017

An Exploration of the Effects of Sensory Stimuli on the Completion of Security Tasks

Bruce Berg
UC Irvine
bgberg@uci.edu

Tyler Kaczmarek
UC Irvine
tkaczmar@uci.edu

Alfred Kobsa
UC Irvine
kobsa@uci.edu

Gene Tsudik
UC Irvine
gts@ics.uci.edu

Abstract— The number and variety of security-critical tasks requiring human involvement has been growing. Such tasks are designed to minimize errors and maximize task performance. It is assumed that task complexity is the main reason for errors. However, ambient sensory distractions might also play a role. These effects have been largely unexplored. It is unclear whether adversarial control over human sensory input can broaden the attack surface. To shed some light on this issue, we conducted large-scale experiments that exposed subjects to unexpected audio and visual stimuli while they performed a security-critical task. Results show that distinct stimuli yield different effects on task performance. In general, less complex stimuli improve subject performance, while more complex stimuli worsen it. This study was conducted in an automated and unattended experimental setting. We discuss our experience, including the potential for abuse of overstimulation as well as benefits and limitations of the unattended experimental paradigm.

Keywords—Usability; Usable Security; Security-Critical Tasks; Distractions; Sensory Stimulation; Bluetooth Pairing

I. Introduction [1]

Both the number and variety of online services and gadgets grows constantly. This results in commensurate growth (also in the number and variety) of security-critical tasks that require human involvement. Commonplace examples include: (1) entering a password or PIN, (2) copying and entering a one-time token as second-factor authentication, (3) solving a CAPTCHA, (4) comparing numbers while pairing Bluetooth devices, (5) using biometric devices, and (6) answering personal security questions.

Since overall security of such tasks is determined by the human user (as the weakest link), many usability studies have been conducted to assess users' ability to perform such tasks correctly and quickly while providing an acceptable level of security [4, 16, 8, 9, 10, 19]. However, since these studies are usually conducted in sterile lab-like environments, they do not reflect typical real-world use-cases. Specifically, they do not take into account the effects of unexpected sensory stimuli, which could be used as an attack vector in adversary-controlled environments.

Security-critical tasks that require human participation are specifically designed to minimize human errors. Therefore, trials with numerous subjects are needed to collect sufficient data in order to accurately assess human error rates. This is exacerbated by the need to test multiple task modalities, each with a distinct set of subjects. To lower this logistical burden, we designed an entirely unattended and fully automated experimental setup. In it, a subject receives pre-recorded instructions from a life-sized projection of an experimenter ("avatar") and has no interactions with a live experimenter.

In this setting, we experimented with about 300 subjects who attempted to pair two Bluetooth devices (one of which was their own) in the presence of various unexpected stimuli. We anticipated that introduction of such stimuli would have negative effects on subject task completion. Surprisingly, we discovered that it did not have a uniform impact on performance: we observed both positive, negative, and no effects, depending on the type of stimulus. This gives rise to a broad range of potential security interventions. Using carefully selected stimuli, an adversary can cause subjects to fail, while a benefactor can improve their success rates. The rest of the paper is organized as follows: The next section overviews related work and background material. We then present the design and setup of the experiments, followed by the presentation and discussion of our results. Next, we discuss the implications and summarize lessons learned. The paper concludes with the discussion of future work.

II. Background & Related Work

This section overviews related work in automated experiments and provides background information from psychology, particularly, on effects of sensory arousal on task performance.

A. Automated Experiments

We are unaware of any prior large-scale usability studies utilizing a fully automated and unattended physical environment. However, some prior work confirms the validity of virtually-attended remote experiments and unattended online surveys in comparison to similar efforts in a traditional (attended) lab-based setting. For example, Ollesch et al.[14] collected psychometric data in: a physically attended

[1] Portions of this work have appeared in [7] and [2].

experimental lab setting, and its virtually attended remote counterpart. No significant differences between the two sets were found. This is further reinforced by Riva et al.[18] who compared data collected from unattended online and attended online, questionnaires. Finally, Lazem and Gracanin [11] replicated two classical social psychology experiments where both the participants and the experimenter were represented by avatars in Second Life[2], instead of being physically co-present. Here too, no significant differences were observed. While these prior experimental settings are not exact analogs to our setup, they indicate that the automated and unattended nature of our experiment should not affect its validity.

B. *Effects of Sensory Stimulation*

Sensory stimulation has variable impact on task performance. This is due to many contributing factors, including individuals' current level of arousal. The Yerkes-Dodson Law [3] stipulates an inverse quadratic relationship between arousal and task performance. It implies that across all contributing stimulants, individuals who are either at a very low or very high level of arousal are unlikely to perform well, and that there exists an optimal level of arousal for correct task completion.

An extension to this law is the notion that the completion of simpler tasks – which produce lower levels of initial arousal in subjects – benefits from the inclusion of external stimuli. At the same time, completion of complex tasks which produce a high level of initial arousal suffers from inclusion of external stimuli. Hockey [6] and Benignus et al. [1] classified this causal relationship by defining the complexity of a task as a function of the task's event rate and the number of sources that originate these subtasks. External stimulation can serve to sharpen the focus of a subject at a low arousal level, thereby improving task performance as found by Olmedo [16]. Conversely, Harris found that stimulation can overload subjects who are already at a high level of arousal, and induce errors in task completion [5].

III. MEHODOLOGY

This section describes our experimental setup, procedures and subject parameters.

A. *Physical Setting*

The experimental setting was designed to facilitate fully automated experiments with a wide range of sensory inputs in a semi-public setting. While we wished to avoid the contrived and unrealistically sterile confines of a traditional lab, we also needed to avoid sporadic interference due to passersby. Consequently, we picked a low-traffic (though publicly accessible) alcove at the top floor of a 6-story building that houses an Information and Computer Sciences school of a large public university. Figure 1 shows the setup from the subject's perspective and from the side. The setup is entirely comprised of readily available off-the-shelf components:

- A 60"-by-45" touch-sensitive interactive Smartboard whiteboard with a Hitachi CP-A300N short-throw projector. The Smartboard acts as both an input and a display device. It reacts to tactile input, similar to a large touch-screen[2].
- An iMAC that uses the SmartBoard as an external display and also serves as the opposite Bluetooth device for the pairing process. The iMAC is hidden from subject's view; it is located directly on the other side of the SmartBoard wall in a separate office.
- A Logitech C9220 HD Webcam[23].
- Two pairs of BIC America RtR V44-2 speakers: one alongside the smartboard, and the other on the opposite wall. Their arrangement is such that the subject will typically stand in the center of the four speakers[2].
- Four programmable wirelessly controllable Phillips Hue A19 LED lightbulbs to deliver the visual stimuli[2].

The final component was the subject's personal Bluetooth-capable device. All recruitment materials stated that prospective subjects were required to bring such a device. Alternatively, we could have provided subjects with a Bluetooth-capable device, which would have streamlined subjects' experience. However, there would have been several drawbacks:

- We did not want to introduce additional errors due to subjects' unfamiliarity with the pairing device. Additional training would be needed to mitigate such errors, which would be infeasible within the unattended paradigm..
- Almost all real-world Bluetooth pairing scenarios involve a user-owned device. Introduction of an experimenter-owned device would reduce external validity of the study.
- Unlike other equipment (which was bulky and physically attached to surfaces), a mobile device could be theft-prone in an unattended setting.

Unsurprisingly, the vast majority of subjects' devices (270 out of 296) were smartphones. The rest were tablets (20) and laptops (6).

[2]See: secondlife.com

[3] See: meethue.com for Hue Bulbs, smarttech.com for the Smartboard, logitech.com for the Webcam, bicamerica.com for speakers, and hitachi.com for the projector.

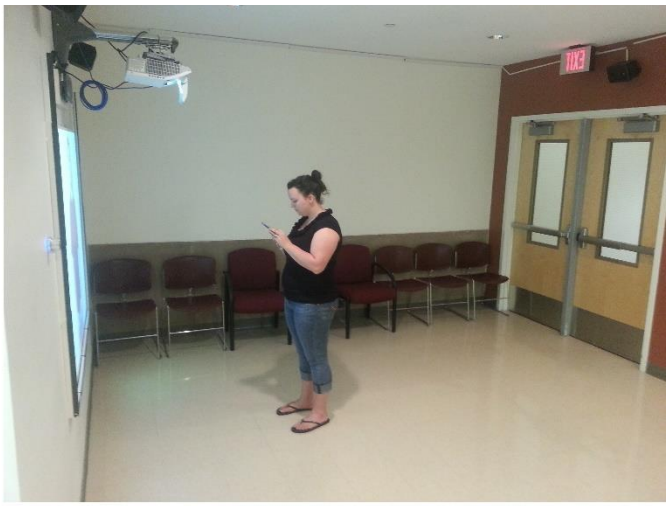

Figure 1: Experiment Setup: Subject's Perspective and Side View

B. Bluetooth Pairing Rationale

Bluetooth pairing is not as common as other security-critical tasks, such as entering passwords, answering security questions, or solving CAPTCHAs. Nevertheless, it is an excellent task for empirical user studies. The security-critical component in Bluetooth paring is comparison of two short strings of about 6 digits. The strings are presented to the user on displays of both devices and the user must confirm whether they match. This requires a single button-press, making Bluetooth pairing a uniform task of unvarying difficulty. It avoids some pitfalls of PIN/password entry, or answering security questions, since no secrets are involved. Meanwhile, other tasks (such as CAPTCHA solving) have widely varying difficulty levels. Whereas, Bluetooth pairing should yield more stable error rates, while not requiring users to divulge any secrets.

C. Procedures

As mentioned earlier, no experimenters were present (either physically or virtually) during experiments. Subjects interacted with a life-sized experimenter avatar which performed all subject briefings and was the subjects' sole source of information throughout the experiment. All subject recruitment materials were deliberately vague and mentioned only the usability focus of the experiment. Actual experimenter involvement was limited to strictly off-line activities, such as: infrequent recalibration of avatar video volume, stimulus volume, and visual effects, as well as occasional repair of components that suffered minor wear-and-tear. The unattended setup allowed the experiment to run continuously, for long periods of time. Specifically, it was conducted over several months-long intervals throughout 2014 and 2015.

The experiment ran in four phases:

1. **Initiation**: Subject initiates the experiment by pressing a large silver button next to the SmartBoard. Duration: instant.
2. **Instruction**: Instructions are given by the avatar. Duration: 45 seconds..
3. **Task Completion**: Subject attempts to pair their personal device with our remote device. Subject is exposed to one randomly selected auditory or visual stimulus, administered through four overhead speakers or light bulbs. Duration: up to 3 minutes..
4. **Compensation**: Subject is asked to fill out a brief demographic survey and enter an email address on the touch-sensitive SmartBoard to receive compensation (a $5 Amazon gift card). Duration: up to 6 minutes.

The total duration ranges between 5 and 10 minutes.

Avatar instructions described the nature of the experiment and the task. Subjects are informed before the 15 second mark that the task requires using the Bluetooth feature of their personal wireless device, thus leaving over 30 seconds to enable Bluetooth Discovery Mode, if not already enabled.

The task completion phase included three stages, with a different stimulus in each. The first stage comprised naturally occurring, static audio stimuli. The second included static and dynamic visual stimuli. The third was limited to one fabricated dynamic looming sound.

Each static audio stimulus was played at constant volume during the entire three minute pairing window, in equal balance from the speakers located above and behind the subject. We used the following four static sounds (volumes measured at the typical subject position):

1. Crying Baby: 67 dB
2. Helicopter: 79 dB
3. Hammer: 80 dB
4. Saw: 78 dB

The volume of the dynamic looming stimulus increased from nearly silent to 85 dB over 5 seconds. After the sound ended, it repeated at a different left/right and front/back

speaker balance, selected randomly. This repeated continuously during the entire three-minute pairing window. Even the highest of these volumes (85 dB) is well within the safe range of the US Occupational Safety & Health Administration (OSHA) guidelines.[4].

For the second set of experiments, we selected six visual effects that differed along two dimensions: color and intensity. Color conditions were picked based on capabilities of programmable light bulbs as well as background knowledge about emotive effects of color. Phillips Hue is an LED system that is based on creating white light. It can produce neither a blacklight effect nor any achromatic light, which limits color selection to the subspace of the CIE color space [20] that Hue supports.

With that restriction in mind, we looked into research on emotive reception and sensory effects in the Munsell color space [15]. It showed that principal hues -- Red, Yellow, Purple, Blue, and Green -- are typically positively received. In contrast, intermediate hues (i.e., mixtures of any two principal hues) are more often negatively associated. It also demonstrated that exposure to different colors can yield either an arousing or a relaxing effect on a subject [13]. Armed with this information, we chose three colors that differ as much as possible, in order to maximally diversify stimulus conditions:

Red: Principal hue with positive emotional connotations, high associated arousal levels

Blue: Principal hue with positive emotional connotations, low associated arousal levels

Yellow-Green: Intermediate hue with negative emotional connotation, high associated arousal levels

We selected two intensity conditions and applied them to all color choices. A more complex modality is generally more arousing and has a greater effect than its simpler counterpart [12]. Since we could not find any previous work on impacts of exposure to colored light on performance of security-critical tasks, we decided to include the simplest modality of exposure that corresponds to the lowest possible level of induced stimulation as the first intensity type. A second, more complex, modality was included to observe the effect of conditions of varying complexity.

The first intensity condition was *Solid*, wherein Hue bulbs were set to constant maximum intensity for the duration of the pairing window. The second was *Flickering*, wherein intensity waxed and waned from Hue bulb's maximum to minimum and back, cycling with a frequency of 1Hz for the duration of the pairing window. In all 6 settings, we used maximum saturation. CIE Color parameters [20] for Phillips Hue bulbs in three color conditions and two intensity conditions are:

1. Red, CIE Chromatic Value: X= 0.674, Y = 0.322
2. Blue, CIE Chromatic Value: X = 0.168, Y = 0.041
3. Yellow-Green, CIE Chromatic Value: X = 0.408, Y =0.517
4. Solid intensity lumen output: 600 lm
5. Flickering intensity lumen range: 6 lm - 600 lm

Our choice of intensity conditions is not unique. For example, we could have included a more complex and startling *Strobing* condition, achievable through rapid modulation of light intensity. It probably would have had a more pro- found impact on the subjects. However, ethical and safety considerations, coupled with the unattended nature of the experiment, precluded the use of any condition that could endanger subjects with certain sensitivity conditions, such as photosensitive epilepsy. This led us to select a safe flickering frequency of 1Hz.

We also found that all selected colors (under both intensity conditions) do not interfere with readability of a backlit personal wireless device or the image projected on the SmartBoard. All experimenters, including one who used corrective lenses, could accurately read the screens of their personal devices in all scenarios.

D. Psychophysical Description of Stimuli

The two types of auditory stimuli – real-world and synthetic – have the potential to produce different effects. Selection of real-world sounds (i.e. jack-hammer, baby crying, etc.) are guided by the intent of eliciting a negative emotional response and/or increased level of general arousal. It is reasonable to expect a negative impact of these sounds on task performance. On the other hand, most humans are quite adept at remaining unaffected by the sound of a crying baby during the completion of a critical task. It might even be the case that the urgency conveyed by the sound a crying baby or the potential danger signaled by the sound of a jack-hammer have the effect of sharpening one's focus.

In cognitive sciences, attention is viewed as a limited resource. Any capture of an individual's attention by an aversive stimulus is likely to be momentary, occurring primarily when the stimulus is first introduced. This is because the human attentional system is primed to react quickly to a change in one's environment. This makes evolutionary sense, since a change in the environment can represent an eminent threat. When a change is detected, the attentional system focuses on the source of this change and assesses whether or not it poses a direct threat. Once an assessment is made that a stimulus does not require a response, adaptation to the stimulus from a foreground target into a background context proceeds relatively rapidly as attention is redistributed to other demands. Although an aversive stimulus may remain aversive throughout its presentation, its capacity to disrupt performance of a complex task might rapidly fade.

Synthetic sounds are designed to attract attention resources without necessarily being aversive. To the auditory

[4] OSHA requires all employers to implement a Hearing Conservation Program where workers are exposed to a time-weighted average noise level of 85 dB or higher over an 8-hour work shift. Our noise levels were clearly lower. See: https://www.osha.gov/SLTC/noisehearingconservation/

attention system, a looming sound could embody a context of constant change, essentially "tricking" the system into a state of sustained engagement. We expect that synthetic sounds have a greater and more sustained effect than their natural counterparts.

E. Recruitment

The main challenge encountered in the recruitment process was scale. Prior usability studies of human-aided pairing protocols [4, 17, 8, 19] demonstrated that 20-25 subjects per tested condition represents an acceptable size for obtaining statistically significant findings. the experiment has one condition for each of the five auditory, as well as one condition for each of the six visual stimuli variations, plus the control condition with no distractions. Therefore, collecting a meaningful amount of data requires at least 240 experimental runs..

We used a four-pronged strategy to recruit subjects:

1. Email announcements sent to both graduate and undergraduate students.
2. Signboards near the entrance and in the lobby of the campus building that housed the experimental setup.
3. Several instructors promoted the experiment in their classes.
4. Printed fliers handed out by experimenters at various campus locations during daily peak pedestrian traffic times.

All recruitment materials announced that subjects were sought for a brief "Usability Study" and that they needed to have a personal Bluetooth-capable device. No mention was made of the security-critical nature of the task, nor the possibility of any kind of stimuli. The materials directed prospective participants to the building in the Computer Science and Engineering quadrant of campus that houses the experimental setup, and mentioned the Amazon gift card reward.

Recruitment efforts yielded 296 subjects of which three quarters were male. This is expected given the campus location of the experimental setup. The overwhelming majority of the subjects (276) were of college age (18-24 years), while 14 were older (25+.) This distribution is not surprising given the typical university population and the fact that being an experimental subject is much less attractive for the older population that generally consists of researchers, faculty and staff. Subjects' demographics were thus heavily geared towards young, tech-savvy male undergraduates.

F. Data Cleaning

There were three reasons for discarding experimental data:

First, although recruitment materials explicitly stated that subjects were to arrive alone and perform the experiment without anyone else present, 37 groups of subjects were observed. However, the nature of the experiment forced subjects to preform the task one-at-a-time, and we found that the initial participant from each group performed in a manner consistent with individual subjects. Subsequent group members, on the other hand, were (unsurprisingly) significantly faster and more successful in their task completion. Consequently, we only considered the data of the first subject in each such group.

Second, some subjects completed the experiment several times, perhaps hoping to receive multiple participation rewards. This occurred even though subjects were explicitly informed that after successfully completing the experiment, all subsequent participation would be discarded.

Third, we discarded otherwise compliant subjects who exhibited obvious visual or auditory impairment. A subject with an auditory impairment would have difficulties understanding the avatar's spoken instructions. A visually impaired subject would experience difficulties in the use of the SmartBoard, and in the pairing process which relies on reading and comparing numbers. However, after carefully reviewing all subject video records we did not identify any obvious visual or auditory impairments.

Finally, in experiments with visual stimuli, each subject was exposed to a single color condition and was not required to distinguish between multiple colors. Because of this, color-blindness should have had minimal impact on results, and we did not vet for it.

IV. Results

This section presents and then discusses experimental results.

A. Task Failure Rate

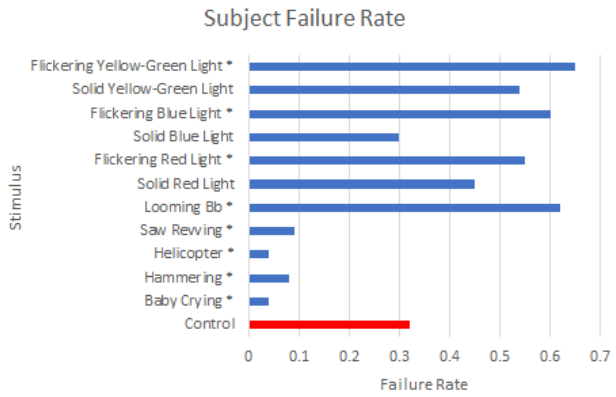

Figure 2: Failure Rates by Stimulus, and Comparison with Control (* = p <0.05 uncorrected)

Figure 2 shows the failure rate for the control condition and each stimulus condition. Applying Barnard's Exact Test pairwise between each stimulus condition and the control condition shows that many differences between failure rates are statistically significant ($p < 0.05$) with respect to all five sound stimuli. However, these stimuli do not impact subject success rates uniformly. We discuss implications of this divergent result in the following section.

B. Task Completion Times

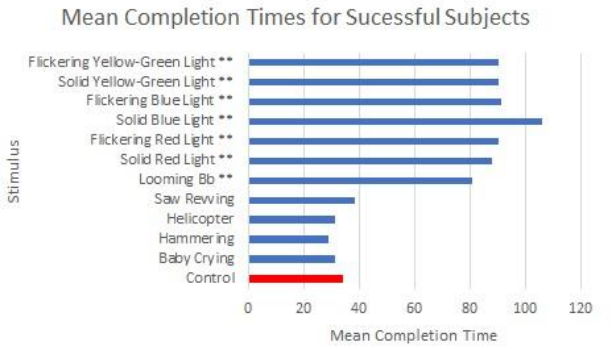

Figure 3: Completion Times for Successful Pairings by Stimulus in Seconds, and Comparison with Control (= $p < 0.001$ uncorrected)**

Figure 3 shows average completion times in successful trials under each stimulus, as well as the results of a one-tailed unpaired t-test on each stimulus condition and the control condition. None of the static audio conditions shows completion times significantly different from control, while other conditions slow down subjects considerably. Next, we examine possible causes for this slowdown.

C. Correction for Multiple Comparisons

To arrive at conclusions in Figures 2 and 3, eleven statistical inferences were needed in each case. The probability of false positives increases with each comparison. Roughly speaking, one can expect one false positive when performing 20 tests with p value of about 0:05. Corrections must be made to maintain the same overall acceptance level for a conjoint outcome. With regard to completion times (Figure 3), we found a statistically significant departure from control for all visual conditions and the dynamic sound condition, even after performing a Holms-Bonferroni correction for 11 comparisons. As far as failure rates (Figure 2), the conjoint outcome is not statistically significant after a Holms-Bonferroni correction for 11 comparisons. We therefore must hedge our conjoint claims. Figure 2 shows that subjects exposed to static auditory stimuli *seem to experience* significant decrease in failure rate, while those exposed to the dynamic audio stimulus and the dynamic visual stimuli *seem to experience* a significant increase.

V. Discussion of Observed Effects

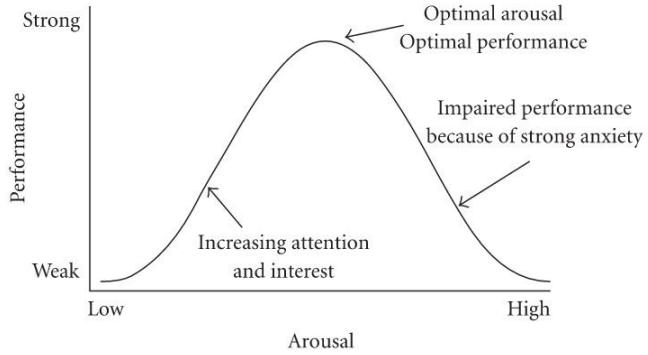

Figure 4: The Yerkes-Dodson Relationship Between Sensory Arousal Levels and Performance

As results show, introduction of unexpected sensory stimuli does not have a uniform effect on subject performance. We found that subjects' error rates go up or down depending on the specific stimulus. Static audio stimuli seem to improve success rates, while dynamic looming audio stimuli and dynamic visual stimuli seem to negatively impact them. Finally, introduction of static visual stimuli has no significant effect on error rates.

The above is consistent with the Yerkes-Dodson Law, which (as mentioned earlier) states that a subject's overall level of sensory arousal is the determining factor in performing any task. When at a low level of arousal, the subject is uninterested and unengaged with the task at hand, thus not performing optimally. Similarly, when overstimulated, the subject is likely to split attention between the arousing stimuli and the task at hand; thus, task performance suffers.

However, there is a middle ground where the overall arousal level allows the subject to be engaged with the task (yet not overwhelmed by it) and yields optimal performance. The relationship between sensory arousal and performance generally follows an inverse U-shaped curve, as Figure 4 illustrates. Considering this curve, we separate the discussion into implications of positive and negative effects.

A. Positive Effects

Intuitively, many subjects were probably not fully engaged when performing Bluetooth pairing. Their general level of sensory arousal during the experiment is analogous to the typical engagement of performing a rote/routine/boring security-critical task. Because of this, introducing a low level of additional sensory arousal can be beneficial to a subject's performance.

Static sound stimuli were the simplest: they served to pique subjects' attentional system and sharpen their focus on the task. At the same time, they did not overload the subjects' attentional system. The fact that only simplest stimuli seem to yield a beneficial effect illustrates the fine line between optimal sensory arousal and over-stimulation. However, this beneficial effect opens up the potential for benevolent actors to include such simple sensory stimulation during security-

critical tasks in order to push subjects along the Yerkes-Dodson curve towards the optimal level of sensory arousal and thereby towards optimal performance.

B. Negative Effects

All dynamic stimuli had a significant negative impact on subjects' completion rates. Consequently, with both visual and auditory stimulation, it is the dynamism of a stimulus (and not its emotional connotation) that determines the level of sensory arousal that a subject experiences.

Negative effects on success rates could motivate an adversary who controls light or sound in the physical setting of a security-critical task. By using a highly dynamic stimulus, an adversary could conduct a denial-of-service (DoS) attack by inducing user failure. Since the emotional connotation of tested stimuli did not have any impact on subject performance, such an attack could be made even more insidious using only positively-perceived stimuli.

While there is a potential to attack subjects performing Bluetooth pairing, a much greater impact was observed in terms of completion times. Subjects would often avert their gaze from their personal device immediately upon exposure to a stimulus. Subjects would then typically glance in the direction of the source of the stimulus (i.e., speakers or lights) and then return their gaze to the personal device. The resulting delay frequently caused the subject's device to exit the Bluetooth pairing menu due to a time-out. The subject would then have to re-start the Bluetooth pairing protocol, resulting in a much longer completion time.

This effect can pave the way for attacks, as discussed earlier. One possibility is that the adversary's goal is DoS, i.e., it aims to bungle users' pairing attempts through added delay. In another scenario, the adversary would try to "buy time" via introduction of sensory stimuli, while interposing its own malicious device(s) and then attempt to fool the user into pairing with that device. In the worst case, the adversary might take advantage of users' inattentiveness and trick them into accepting a non-matching authenticator.

C. Unattended & Automated Setup

We believe that the fully unattended/automated experimental paradigm is advantageous and applicable to many other settings. There will always be a certain logistical burden in continuously running an experiment for months at a time. Our setup offers two unique advantages over its traditional attended counterpart:

1. **Impromptu participation**: a subject just shows up at will and participates in the experiment. There is no need to pre-schedule time-slots by email, on the web or in person. This significantly lowers the participation barrier.
2. **No human presence**: no human attendant is needed to facilitate correct flow of the experiment. This is in contrast to expecting one or more people to be constantly present (for hours on end), which results in greater logistical and financial overheads.

In the unattended setup, all timing and completion data was collected automatically. Off-line review of all subject video recordings was rather superficial. The goal was only to confirm subjects' compliance with instructions and it translated into several seconds per subject. We believe that this setup saved approximately 100 man-hours of human attendant's time over the traditional attended setup. These savings are two-fold: (1) time spent scheduling time-slots and administering experiment sessions, and (2) time spent waiting for subjects to arrive. Finally, our use of a single instruction set given in uniform manner to each subject resulted in reduced variance.

VI. Conclusions and Future Work

As human participation in security-critical tasks becomes more commonplace, users are more likely to attempt these tasks in environments where they could be exposed to potentially malicious sensory distractions. This trend motivates studying the impact of external stimuli. Research described in this paper sheds some light on the relationship between completion of security-critical tasks and exposure to unexpected stimuli. However, this work is only the beginning.

Given the observed negative effect on subject completion times for complex stimuli, one interesting next step is to conduct a similar experiment where subjects are frequently shown non-matching codes during the Bluetooth pairing process. In this setup, a subject's acceptance of non-matching codes as matching would represent a successful attack by an adversary seeking to pair the subject's device with a malicious device. This experiment would thus focus on effects of sensory stimuli on successful deception rates.

Furthermore, we plan to conduct a study of subjects performing security-critical tasks while exposed to multiple auditory stimuli lasting longer than 3 minutes. This might allow us to learn whether subjects' sensory arousal is the result of (1) "startling" the human attentional system with a sudden unexpected stimulus, or (2) an unavoidable psycho-physical reaction.

Finally, we might try to physically measure subject arousal with an electroencephalogram (EEG) to gain a more precise understanding of subject arousal levels during task completion. However, this would be costly and incompatible with the unattended experimental paradigm.

VII. Acknowledgements

This research was supported by NSF grant CNS-1544373. We thank the anonymous journal reviewers for their constructive comments.

VIII. References

[1] V. A. Benignus, D. A. Otto, and J. H. Knelson. Effect of low-frequency random noises on performance of a numeric monitoring task. Perceptual and motor skills, 40(1):231-239, 1975.

[2] B. Berg, T. Kaczmarek, A. Kobsa, and G. Tsudik. Lights, camera, action! Exploring effects of visual distraction on completion of security tasks. arXiv:1705:xxxx

[3] R. A. Cohen. Yerkes-Dodson law. In Encyclopedia of clinical neuropsychology, pages 2737-2738. Springer, 2011.

[4] A. Gallego, N. Saxena, and J. Voris. Exploring extrinsic motivation for better security: A usability study of scoring-enhanced device pairing.

In A.-R. Sadeghi, editor, Financial Cryptography and Data Security, volume 7859 of Lecture Notes in Computer Science, pages 60–68. Springer Berlin Heidelberg, 2013.

[5] W. Harris. Stress and Perception: The Effects of Intense Noise Stimulation and Noxious Stimulation upon Perceptual Performance. Ph.D. thesis, University of Southern California, 1960.

[6] G. R. .J. Hockey. Effect of loud noise on attentional selectivity. The Quarterly Journal of Experimental Psychology, 22(1):28-36, 1970.

[7] T. Kaczmarek, A. Kobsa, R. Sy, and G. Tsudik. An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise. In *Proceedings of the NDSS Workshop on Useable Security 2015*, pages 14:1-14:12

[8] R. Kainda, I. Flechais, and A. W. Roscoe. Usability and security of out-of-band channels in secure device pairing protocols. Proceedings of the 5th Symposium on Usable Privacy and Security, pages 11:1-11:12, 2009. ACM ID: 1572547.

[9] R. Kainda, I. Flechais, and A. W. Roscoe. Two heads are better than one: security and usability of device associations in group scenarios. In Proceedings of the Sixth Symposium on Usable Privacy and Security, SOUPS '10, pages 5:1-5:13, 2010. ACM ID: 1837117.

[10] S. Laur, N. Asokan, and K. Nyberg. Efficient mutual data authentication using manually authenticated strings. Cryptology ePrint Archive, Report 2005/424, 2005. http: //eprint.iacr.org/

[11] S. Lazem and D. Gracanin. Social traps in second life. In 2010 Second International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), pages 133-140, Mar. 2010

[12] H. S. Koelega, J.-A. Brinkman, B. Zwep, and M. N. Verbaten. Dynamic vs static stimuli in their effect on visual vigilance performance. Perceptual and motor skills, 70(3):823-831, 1990.

[13] K. Naz and H. Epps. Relationship between color andemotion: A study of college students. College Student J, 38(3):396, 2004.

[14] H. Ollesch, E. Heineken, and F. P. Schulte. Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. International Journal of Internet Science, 1(1):71-81, 2006.

[15] D. Nickerson. History of the munsell color system and its scientific application. Journal of the OpticalSociety, 1940.

[16] E. L. Olmedo and R. E. Kirk. Maintenance of vigilance by non-task-related stimulation in the monitoring environment. Perceptual and motor skills, 44(3):715-723, 1977.

[17] E. Uzun, K. Karvonen, and N. Asokan. Usability analysis of secure pairing methods. In S. Dietrich and R. Dhamija, editors, Financial Cryptography and Data Security, volume 4886 of Lecture Notes in Computer Science, pages 307-324. Springer Berlin Heidelberg, 2007.

[18] G. Riva, T. Teruzzi, and L. Anolli. The use of the Internet in psychological research: comparison of online and offine questionnaires. CyberPsychology & Behavior, 6(1):73-80, 2003.

[19] C. Paul, E. Morse, A. Zhang, Y.-Y. Choong, and M. Theofanos. A field study of user behavior and perceptions in smartcard authentication. In Human-Computer Interaction, INTERACT 2011, volume 6949 of LNCS, pages 1-17. Springer Berlin / Heidelberg, 2011.

[20] G. Wyszecki and W. S. Stiles. Color science, volume 8. Wiley New York, 1982.

Author Bios

Bruce Berg (bgberg@uci.edu)

Bruce G. Berg is an Associate Professor in the Department of Cognitive Sciences at the University of California, Irvine. His research interests are in auditory attention and perception, theoretical psychoacoustics, and signal detection theory. Early in his career, he originated the technique of adding noise to stimuli as a means for investigating attention. Current work includes the development of a theory in which the filtering properties of the auditory periphery are different for spectral and temporal processes. He received a Ph.D. from Indiana University in Psychology and was awarded a NIH Postdoctoral Fellowship from Brigham and Women's Hospital where he used signal detection theory to investigate the strategies of radiologists in reading images.

Tyler Kaczmarek (tkaczmar@uci.edu)

Tyler Kaczmarek is a fourth-year Ph.D candidate at the Donald Bren School of Information and Computer Science of the University of California, Irvine currently studying under Professor Gene Tsudik. Tyler's principal areas of interest are useable security and biometric techniques for continuous authentication.

Alfred Kobsa (kobsa@uci.edu)

Alfred Kobsa is a Professor in the Donald Bren School of Information and Computer Sciences of the University of California, Irvine. He received his master degrees in Computer Science and in the Social and Economic Sciences from the Johannes Kepler University Linz, Austria, and his Ph.D. in Computer Science from the University of Vienna, Austria and the Vienna University of Technology. Dr Kobsa's research lies in the areas of personalized systems, privacy, usable security, and support for personal health maintenance. He was the founding editor of User Modeling and User-Adapted Interaction: The Journal of Personalization Research, and the founding president of User Modeling Inc. Dr. Kobsa edited several books and authored numerous publications in the areas of user-adaptive systems, privacy, human-computer interaction and knowledge representation.

Gene Tsudik (gts@ics.uci.edu)

Gene Tsudik is a Chancellor's Professor of Computer Science at the University of California, Irvine (UCI). He obtained his PhD in Computer Science from USC in 1991. Before coming to UCI in 2000, he was at IBM Zurich Research Laboratory (1991-1996) and USC/ISI (1996-2000). His research interests included numerous topics in security and applied cryptography. He currently serves as Director of Secure Computing and Networking Center (SCONCE) at UCI. Gene Tsudik is a former Fulbright Scholar and Fulbright Specialist, a fellow of ACM, IEEE and AAAS, as well as a foreign member of Academia Europaea. From 2009 to 2015 he was the Editor-in-Chief of ACM Transactions on Information and Systems Security (TISSEC).